最大エントロピーと最小ダイバージェンスの双対性

江口　真透　　数理推論研究系

概要 ○ 生態学、計算言語学において**最大エントロピー法**が広く使われている。特徴変数から統計モデルと同時に統計推定が自 動的に得られるのでデータ解析が容易に行うことができる。

○ パッケージMaxEnt (Phillip *et al.* 2006) を使えば、ある生物種の在データと、その調査エリアのGIS データ、気象データなどのリンクからの特徴変数データが構成され、Lassoタイプの変数選択による**生息マップ**の描画が簡単に得られる。

○ 統計学の見地からは、この最大エントロピー法は統計モデルは**指数型分布族**、推定は**最尤法**に限られているのでモデルの誤特定とデータの混濁に対してロバストでない点が挙げられる。

○ この問題を解決する方法は幾つか考えられるが、ここでは**エントロピーの一般化**による最大化を提案したい。このアプローチは生 態学では生物多様性を測る別のエントロピーが考えられていることからも支持される。

○ 一般化最大エントロピーモデルに対して**最小ダイバージェンス法**を適用すると自然な振る舞いが保持されることを示し、０和ゲームのミニマックス性によって最大エントロピーと最小ダイバージェンスの**双対性**について明らかにする。

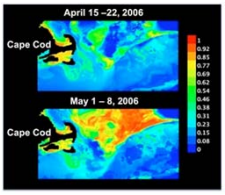
最大エントロピー法とは

ある生物種の調査エリア A の生息分布を推定したい. そのために在データ $\{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$ と同時にGISや気象から特徴変数ベクトル $\{\boldsymbol{t}(\boldsymbol{x}_1),...,\boldsymbol{t}(\boldsymbol{x}_n)\}$ が得られたとする. このとき、統計量 $\bar{\boldsymbol{t}}=n^{-1}\sum \boldsymbol{t}(\boldsymbol{x}_i)$ に対して $\boldsymbol{t}(\boldsymbol{X})$ の平均がちょうど $\bar{\boldsymbol{t}}$ となる分布の集合を考える。この平均の制約の下でエントロピー最大分布は次のようになる。

$$\hat{f} := \underset{f\in\mathscr{F}(\bar{\boldsymbol{t}})}{\arg\max}\Big\{-\sum_{\boldsymbol{x}\in A} f(\boldsymbol{x})\log f(\boldsymbol{x})\Big\} \;\Rightarrow\; \hat{f}(\boldsymbol{x})=\exp\{\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x})-\kappa(\hat{\boldsymbol{\theta}})\},$$

ここで $\mathscr{F}(\bar{\boldsymbol{t}})=\{f:\mathbb{E}_f\{\boldsymbol{t}(\boldsymbol{X})\}=\bar{\boldsymbol{t}}\}$, $\kappa(\boldsymbol{\theta})=\log\sum\exp\{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x})\}$, $\hat{\boldsymbol{\theta}}$ は制約 $\mathbb{E}_{\hat{f}}\{\boldsymbol{t}(\boldsymbol{X})\}=\bar{\boldsymbol{t}}$ から決まる

これより生息分布 $\hat{f}(\boldsymbol{x})$ による生息マップが得られる。



大西洋セミクジラの生息分布
http://www.seascapemodeling.org/

一般化エントロピー

上記の(1) で扱われたエントロピーはBoltzmann-Shannonのものであるが、生態学においては生物多様性の観点からSimpson やHillのエントロピーが広く用いられている。このようにより広いクラスにエントロピーを拡張しよう。

単調増加で凸な関数 $U(s)$ に対して共役な凸関数 $U^*(t):=\max_{s\in\mathbf{R}}\{ts-U(s)\}$ を使って、U-エントロピーを

$$H_U(f)=-\sum U^*(f)=-\sum\{f\xi(f)-U(\xi(f))\},\quad \text{ここで } \xi(t)=(U')^{-1}(t).$$

と定義する。　$M_U=\{U'(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x})-\kappa(\boldsymbol{\theta}):\boldsymbol{\theta}\in\Theta\}$, $\Theta=\{\boldsymbol{\theta}:\kappa(\boldsymbol{\theta})<\infty\}$として, M_U を U-モデルと呼ぶ　このとき、

$$\hat{f}_U := \underset{f\in\mathscr{F}(\bar{\boldsymbol{t}})}{\arg\max}\, H_U(f) \;\Rightarrow\; \hat{f}_U(\boldsymbol{x})=U'(\hat{\boldsymbol{\theta}}_U{}^{\mathrm{T}}\boldsymbol{t}(\boldsymbol{x})-\kappa(\hat{\boldsymbol{\theta}})),\ (\kappa(\boldsymbol{\theta})\text{ は正規化項, }\hat{\boldsymbol{\theta}}_U\text{ は平均制約 }\mathbb{E}_{\hat{f}}\{\boldsymbol{t}(\boldsymbol{X})\}=\bar{\boldsymbol{t}}\text{ から決まる})$$

実際、 $f\in\mathscr{F}(\bar{\boldsymbol{t}}) \;\Rightarrow\; H_U(\hat{f}_U)=C_U(f,\hat{f}_U) \iff H_U(\hat{f}_U)-H_U(f)=D_U(f,\hat{f}_U)\geq 0$である.

最小ダイバージェンス

U-クロス エントロピーと U-ダイバージェンスを次のように定める。

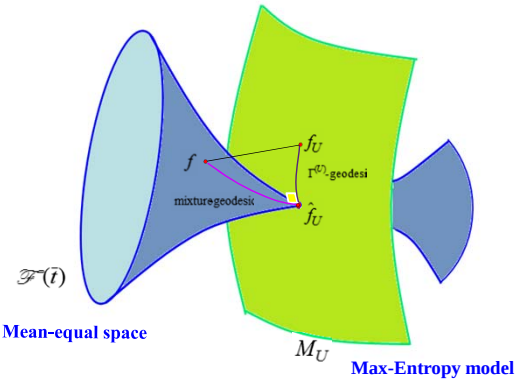
$$C_U(f,g)=\sum\{f\xi(g)-U(\xi(g)),\quad D_U(f,g)=C_U(f,g)-H_U(f)$$

このとき、最小 U-ダイバージェンスと 0 和ゲームの**ミニマックス**性:

$$\hat{\boldsymbol{\theta}}_U \;=\; \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\ \{-\boldsymbol{\theta}^{\mathrm{T}}\bar{\boldsymbol{t}}+\kappa(\boldsymbol{\theta})+\sum U(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{t}+\kappa(\boldsymbol{\theta}))\}$$

$$\max_{f\in\mathscr{F}(\bar{\boldsymbol{t}})}\min_{g\in\mathscr{F}} C_U(f,g)=H_U(\hat{f}_U)=\min_{g\in\mathscr{F}}\max_{f\in\mathscr{F}(\bar{\boldsymbol{t}})} C_U(f,g)$$

が成立する。



これからやりたいこと

- 標準的な最大エントロピー法の欠点を一般化最大エントロピー法によって**改善**したい。このためには生成関数 U をデータから有効に選択する方法の開発が必要となる。一つのアイデアとしては 幾つかの U によって得られた生息分布の性能をテストデータによる検証で選択する方法が考えられる。
- 拡張されたエントロピー最大分布は標準の場合の**ギブス分布**とは異なる。この分布の統計的、生態学的な適切な解釈を得たい。同時に在のみデータの**ポアソン過程**としての理解も拡大したい。